SUPPLEMENTARY MATERIAL:

Supplementary Methods, 7 Figures and one Table (separate .xls file)

HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range *MYB* enhancers

Ralph Stadhouders[1]*, Suleyman Aktuna[2]*, Supat Thongjuea[3,4], Ali Aghajanirefah[1], Farzin Pourfarzad[1], Wilfred van IJcken[5], Boris Lenhard[3,6], Helen Rooks[2], Steve Best[2], Stephan Menzel[2], Frank Grosveld[1,7], Swee Lay Thein[2,8#] and Eric Soler[1,7,9#]

[1]Department of Cell Biology, Erasmus Medical Centre, Rotterdam, the Netherlands
[2]King's College London, Department of Molecular Haematology, London, United Kingdom
[3]Computational Biology Unit, Bergen Center for Computational Science, Bergen, Norway
[4]MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, UK
[5]Centre for Biomics, Erasmus Medical Centre, Rotterdam, the Netherlands
[6]Department of Molecular Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London; and MRC Clinical Sciences Centre, London, UK
[7]Cancer Genomics Center, Erasmus Medical Center, Rotterdam, The Netherlands
[8]King's College Hospital Foundation Trust, Denmark Hill, London, UK
[9]INSERM UMR967, CEA/DSV/iRCM, Fontenay-aux-Roses, France

*equal contribution
[#]equal contribution & corresponding authors

SUPPLEMENTARY METHODS

ENCODE and expression microarray data mining

ENCODE project genome-wide datasets (hg18) were accessed via the UCSC Genome Browser (http://genome.ucsc.edu). ENCODE Integrated Regulation Tracks were used to obtain RNA-seq and DNAseI-seq data, which were combined with Chromatin State Segmentation data (1) (Combined weak and strong enhancer signatures, Broad Institute). Genomic footprinting data was obtained from the UW DNase DGF tracks (University of Washington). Additional histone modification data (H3K27Me3, H3K36Me3) shown in Supplementary Figure 1 was obtained from the Broad Histone tracks (Broad Institute). The following tracks were used for the assessment of erythroid enhancer potential (in K562 cells) of the intergenic LDB1 binding sites (Figure 2F and Supplementary Figure 2): H3K4Me1, H3K4Me2, H3K4Me3 and H3K27Ac (Broad Histone tracks, Broad Institute); p300 (sc48343 IgG-rab, SYDH TFBS tracks, ENCODE/Stanford/Yale/USC/Harvard) and DNaseI-seq (as described above). c-Myb ChIP-seq data was also obtained from ENCODE (mm9 build, Stanford/Yale TFBS track). Mouse-human sequence conservation was determined for the 100 bp centre of each intergenic LDB1 binding site using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start).

Microarray expression analysis datasets of *MYB*-depleted human erythroid progenitor cells were obtained from Sankaran et al. (2) (GSE25678) and Bianchi et al. (3) (GSE13110). Data were analyzed with GEO2R (4) using standard analysis settings to generate lists containing the top 250 affected genes.

Intergenic SNP selection and transcription factor motif prediction

HBS1L-MYB intergenic common DNA variants associated with erythroid traits (Figure 1) were identified from published data (Table 1). More detailed analysis was performed with a distinct group of 17 genetic variants that show peak association across at least three of the main traits (%HbF/%F cells, MCV, MCH and RBC parameters) and across several studies. As an example, a cut-off of $p<10^{-75}$ for MCV was chosen for data from van der Harst *et al* (5). These variants belong to a tight linkage disequilibrium (LD) block (HMIP-2 (6)) detected in Europeans and Asians, spanning ~24 kb (chr6:135,452,921-135,477,194, hg18 co-ordinates). 15 of the 17 variants studied are SNPs (single-nucleotide polymorphisms): rs9376090, rs7776054, rs9399137, rs9389268, rs11759553, rs9373124, rs4895440, rs4895441, rs9376092, rs9389269, rs9402686, rs6920211, rs9494142, rs9494145 and rs9483788.

Two additional variants belonging to this group (a 3bp deletion (7), rs66650371, also known as rs67449035, and a SNP residing in its non-deleted allele, rs7775698) have usually been reported jointly (with strong association) under ‘rs7775698’, since commercial genotyping arrays cannot distinguish between deletion and SNP (7). To assess whether these ‘candidate variants’ reside within or near predicted GATA1, TAL1 and/or CTCF binding motifs we used the JASPAR (8) database with profile score thresholds of >75%.

Luciferase reporter assays

The -84 and -71 regulatory region were PCR amplified from K562 genomic DNA (primers: -84F – ACTCTGGACAGCAGATGTTACTAT and -84R – TGAGGGAACCGCCCT; -71F - GTAGTCTAGTATGTATTGGGTTCC and -71R – AAGATCGCGCCACTGCA) and cloned 3’ of the luciferase gene in pGL3-promoter (Promega). Sequence identities of the inserted regions (including the rs66650371 allelic identity of the -84 region) were verified using standard Sanger-sequencing. MEL or HEK cells ($2x10^5$ cells per well) were transfected in a 24-well plate using Lipofectamine LTX (Invitrogen) according to the manufacturer’s instructions. 0.25 µg of pGL3-promoter plasmid was transfected; a TK-Renilla plasmid was cotransfected for normalization purposes. Luciferase activity was measured 48h post-transfection using the Dual-Luciferase Reporter Assay System (Promega) and normalized for renilla levels. Luciferase levels generated by a pGL3-promoter plasmid without an enhancer region were set to 1.

RNA interference

Lentiviral particles expressing shRNA against human LDB1, TAL1 and KLF1 were obtained from the Sigma TRC 1.0 and TRC 1.5 shRNA libraries (The RNAi Consortium, Sigma). 4-5 shRNA were tested per factor, of which 2 were used for further experiments: LDB1 TRCN0000021784 and TRCN0000021785; TAL1 TRCN0000014690 and TRCN0000014691; KLF1 TRCN0000016276 and TRCN0000016277. A scrambled shRNA (SHC002) was used as a control. Cells were harvested 4 or 5 days after transduction and processed for gene expression analysis as described below.

Gene expression analysis

Total RNA was extracted from K562 cells or primary HEPs using TriPure Isolation Reagent (Roche Diagnostics) according to the manufacturer's instructions. First-strand cDNA synthesis and quantitative real-time PCR was performed as described (9). *ACTB* and *HPRT* expression levels were used for normalization purposes.

3C and 3C-Seq data normalization

3C-qPCR signals were normalized as described before (10-11). To normalize for differences in template loading, we used a proximity-based interaction in the *ERCC3* gene (12) and a known, invariant CTCF-mediated long-range interaction, 600-650kb upstream of *HBB* (13). Differences in PCR primer efficiencies were filtered out by running in parallel identical PCR reactions on a randomly digested and re-ligated BAC DNA sample (covering the entire *MYB* locus: RP11-10409; BACPAC Resources), which was spiked with 200 ng/µl human genomic DNA. Amplification signals for the different primer sets obtained with 3C material where then normalized to the signals obtained with the BAC sample (11).

Allele-specific ChIP(-Seq) analysis

We measured differences in ChIP enrichments between the wildtype and the minor alleles of rs66650371 within the -84 LDB1-complex binding site utilizing the loss of a *Mae*III restriction site (GTNAC, Supplementary Figure 4A) on the minor rs66650371 allele. ChIP material from K562 cells (heterozygous for the rs66650371 SNP (7)) was used for PCR with primers spanning rs66650371 (Supplementary Table 1) to linearly amplify the relevant DNA fragments (both for the ChIP samples and input genomic DNA as a control). The resulting 91bp amplicon was purified and 50 ng of PCR product was digested with *Mae*III (2U, 3h at 55°C). Reactions were separated on a 3% agarose gel and scanned using a Typhoon 9410 Molecular Imager (GE Healthcare). Signal densities of the upper 91bp band and the 2 lower (51bp and 40bp) bands were quantified with ImageQuant 5.2 software (Amersham Biosciences). A rs66650371/wildtype ratio was determined for all ChIP samples; input genomic DNA ratios were used for normalization. To ensure similar digestion efficiencies across the individual samples, identical reactions spiked with 25 ng of plasmid DNA were analyzed in parallel on a 1% agarose gel.

ChIP-sequencing for TAL1 in K562 cells was conducted as described above. Using Bowtie (14), the resulting sequencing reads (36 bp) from TAL1 ChIP and input control samples were mapped against both the hg18 human reference genome and a manually modified human genome in which the rs66650371 TAC sequence was deleted. Only reads strictly informative for the specific alleles were used in the comparison. As the number of reads covering rs66650371 was low in the input control sample (data not shown), we verified the 1:1 allelic ratio observed in the input control material sequencing experiment by cloning a rs66650371-containing PCR amplicon in the pGEM-T easy vector using the pGEM-T easy vector system (Promega). Sequencing of 20 individual clones confirmed equal rs66650371 allelic ratios in K562 chromatin (Figure 5B).

Allele-specific 3C analysis

To quantify allelic differences in chromatin looping between the rs66650371 containing -84 regulatory element and the *MYB* promoter in K562 cells, we amplified a 4.5kb composite 3C fragment from the specific ligation event between the -84 and the *MYB* promoter *Bgl*II fragments (Supplementary Figure 4B). In parallel, a control 4.5kb PCR fragment was amplified from all -84 *Bgl*II fragments within the 3C library. These fragments were purified and used as template for a linear (15-25 cycles) PCR amplification using rs66650371-spanning primers. The rs66650371/wildtype allelic ratio was determined as described above for the allele-specific ChIP analysis. Allelic ratios obtained using the control PCR were set to 1 and used for normalization. As an extra control, genomic DNA was analyzed in parallel.

SNaPshot analysis of allele-specific protein binding and expression

To quantify allelic differences in transcription factor binding at the -71kb regulatory element encompassing rs9494142, a fragment of 158bp including the SNP and GATA-1 binding site was PCR amplified. GATA-1 immunoprecipitated chromatin and input genomic DNA were amplified in parallel. PCR products were purified and used for primer extension with the appropriate extension primer (Supplementary Table 1). SNaPshot reactions (15) were electrophorezed on a DNA Sequencer (ABI 3130, Applied Biosystems) and peak heights of each allele were determined using GeneMarker V2 2.0 demo software (SoftGenetics LLC). Experiments were performed on biological samples from 4 individuals heterozygous for rs9494142 (T/C). For each experiment, the ratio of binding to T/C alleles for input and GATA-1 ChIP were calculated independently. T/C ratios for GATA-1 ChIP samples were then normalized to input T/C ratios.

To measure allele-specific *MYB* expression, we selected the rs210796(A/T) SNP in *MYB* intron 4 as the informative SNP. Healthy unrelated individuals heterozygous for the intergenic variants (HMIP-2) were genotyped for rs210796 and 5 individuals heterozygous for both the intergenic variants and rs210796 were recruited. Five individuals homozygous for the intergenic variants and heterozygous for rs210796 were also recruited as controls. Total RNA was isolated from early erythroid progenitor cells and cDNA was prepared using random hexamer primers. A fragment of 260bp encompassing rs210796 was PCR amplified using genomic DNA and cDNA samples, PCR products were used in SNaPshot reactions as described above. For each experiment, the T/A peak height ratio was calculated for genomic DNA and cDNA samples. T/A ratios for cDNA samples were normalized to genomic DNA T/A ratios.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Genome-wide histone modification, expression and DNaseI hypersensitivity analysis reveals an erythroid/hematopoietic-specific regulatory signature for the human *HBS1L-MYB* intergenic region. (A) Genome-wide ChIP-Seq and RNA-Seq data from the ENCODE consortium is displayed for the human *HBS1L-MYB* intergenic region in 9 different human cell types. Histone 3 lysine 4 trimethylation (K4Me3, marking promoters), lysine 4 monomethylation (K4Me1, marking enhancers), lysine 27 acetylation (K27Ac, marking enhancers) and RNA-Seq expression analysis are shown. **(B)** Histone 3 lysine 36 trimethylation (K36Me3, marking productive transcription elongation) and **(C)** lysine 27 trimethylation (K27Me3, marking Polycomb-repressed regions) ChIP-Seq data for the *HBS1L-MYB* intergenic region in 8-9 different human cell types. **(D)** DNaseI-Seq and Digital Genomic Footprinting data for the *HBS1L-MYB* intergenic region in different human cell types.

Supplementary Figure 2: Colocalization of DNaseI-hypersensitivity, conservation and enhancer-associated histone modifications and proteins with intergenic LDB1-complex binding sites. Colocalization (highlighted by blue shading) of the different enhancer-associated marks (K562 tracks in blue, obtained from the ENCODE consortium) and the individual intergenic LDB1-complex binding sites (numbered by distance to the *MYB* transcription start site, LDB1 ChIP-Seq track in black) from primary human erythroid progenitors. Mammalian conservation (Mammal Cons) is shown in the bottom track. Transcription factor (Ldb1-complex) binding in the corresponding mouse region (9) is denoted below each graph.

Supplementary Figure 3: Chromosome conformation capture analysis of the *HBS1L-MYB* locus reveals long-range interactions between intergenic elements and the *MYB* gene in K562 cells. 3C-qPCR experiments on K562 cells (n=4) using the *MYB* promoter as viewpoint.

The locus is plotted on top, with the different 3C restriction fragments (*Bgl*II) used indicated. Interaction frequencies between 2 fragments within the *ERCC3* locus were used for normalization. Error bars display s.e.m.

Supplementary Figure 4: Strategy used to quantify differences in transcription factor binding and promoter looping to the K562 rs66650371 alleles. (A) *Mae*III digestion read-out method used for allele-specific ChIP and 3C assays. The rs66650371 reference allele contains a *Mae*III site, which is removed by the rs66650371 3bp-deletion allele. A 91bp product spanning rs66650371 was PCR amplified and subjected to *Mae*III digest, resulting in 3 fragments: 88bp (representing the 'SNP allele') and 51/40 bp fragments (together representing the 'WT allele'). Fragments were separated using agarose gel electrophoresis and quantified to determine a SNP/WT ratio. Duplicate samples spiked with pGL3 plasmid were digested in parallel to ensure digestion efficiencies were similar across samples. **(B)** Strategy used for rs66650371 allele-specific 3C in K562 cells. Below a schematic of the locus (top), 2 boxed figures depict the 2 *Bgl*II restriction fragments that form the composite -84/promoter 3C fragment. Three primers were designed that generate 2 PCR amplicons (~4.5kb) encompassing rs66650371. Primers 1+3 can only yield a product when the specific -84/promoter composite fragment is present ('-84+prom. specific PCR'). Primers 1+2 are both located on the -84 fragment and will amplify all -84 fragments ('CTRL PCR'). Amplicons are purified from gel and the *Mae*III digestion read-out approach described in (A) is used to determine allelic ratios. Ratios obtained from the 'CTRL PCR' were used for normalization.

Supplementary Figure 5: Differentiation kinetics of WT/WT and SNP/SNP cultures as assayed by FACS analysis. (A) Representative FACS measurements showing the relative increase of GPA positive (late erythroid cells) cells during phase II culture (day 4-day 13) of primary erythroid progenitors. The percentage GPA positive cells was normalized intra-individually for small differences in the percentage of CD71 positive cells; day 4 measurements were set to 1. Cells were obtained from individuals homozygous for the minor allele of the phenotype-associated SNPs (HMIP-2 LD block variants; SNP/SNP) and wildtype control individuals (WT/WT). **(B)** Percentage of CD14+ monocytes present in WT/WT and SNP/SNP erythroid cells during culture. Representative measurements performed at day 4 and day 11 are depicted.

Supplementary Figure 6: GATA1 ChIP experiments on primary human erythroid cultures harvested at terminal stages of differentiation. (A) Representative results of ChIP-qPCR experiments for GATA1 on chromatin prepared from WT/WT HEPs on day 7 and day 11 of erythroid differentiation. **(B)** Representative results of ChIP-qPCR experiments for GATA1 on chromatin prepared from WT/WT and SNP/SNP (homozygous for the minor allele of the phenotype-associated HMIP-2 LD block variants) HEPs on day 11 of erythroid differentiation. The -84 and -71 regulatory elements were assayed for GATA1 binding, the α-globin hs40 region was used as a positive control. Enrichments were normalized to IgG.

Supplementary Figure 7: Plausible model for *MYB*-mediated repression of HbF levels via cell cycle regulation and transcriptional activation of HbF repressor genes. (A) c-Myb ChIP-seq data (obtained from the ENCODE consortium) from MEL cells showing c-Myb binding to the β-globin locus and HbF repressor genes (*Nr2c2* encodes the TR4 protein). **(B)** c-Myb ChIP-seq data from MEL cells showing c-Myb binding to selected cell cycle regulators. **(C)** Analysis of published c-MYB knockdown studies in human erythroid progenitors. Downregulated HbF repressor genes and a selection of affected cell cycle regulators is shown. **(D)** Dual model of *MYB*-mediated HbF repression. Lower *MYB* levels (as a result of disrupting enhancer variants) can lead to HbF induction via increased premature cell cycle termination (‘indirect’, top part), resulting in the generation of more F-cells and a higher HbF level. Fewer proliferation cycles (‘x2’, indicating cell division) will result in a lower red blood cell count (RBC) and a larger mean cell volume (MCV). Alternatively, lower *MYB* levels could result in a loss of proper transcriptional regulation at the β-globin locus and HbF repressor genes (‘direct’, lower part). Reduced activation by *MYB* of known HbF repressors (e.g. BCL11A, KLF1) or disrupted regulation at the β-globin locus could result in γ-globin gene reactivation and subsequent HbF induction.

References

1. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473(7345):43-49.
2. Sankaran VG, et al. MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13. *Proc Natl Acad Sci U S A*. 2011;108(4):1519-1524.
3. Bianchi E, et al. c-Myb supports erythropoiesis through the transactivation of KLF1 and LMO2 expression. *Blood*. 2010;116(22):e99-110.
4. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue):D991-995.
5. van der Harst P, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*. 2012;492(7429):369-375.
6. Thein SL, et al. Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc Natl Acad Sci U S A*. 2007;104(27):11346-11351.
7. Farrell JJ, et al. A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood*. 2011;117(18):4935-4945.
8. Bryne JC, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*. 2008;36(Database issue):D102-106.
9. Stadhouders R, et al. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO J*. 2012;31(4):986-999.
10. El Kaderi B, Medler S, Ansari A. Analysis of interactions between genomic loci through Chromosome Conformation Capture (3C). *Curr Protoc Cell Biol*. 2012;Chapter 22:Unit22 15.
11. Palstra RJ, et al. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet*. 2003;35(2):190-194.
12. Tolhuis B, et al. Looping and Interaction between Hypersensitive Sites in the Active beta-globin Locus. *Mol Cell*. 2002;10(6):1453-1465.

13. Hou C, Dale R, Dean A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A*. 2010;107(8):3651-3656.

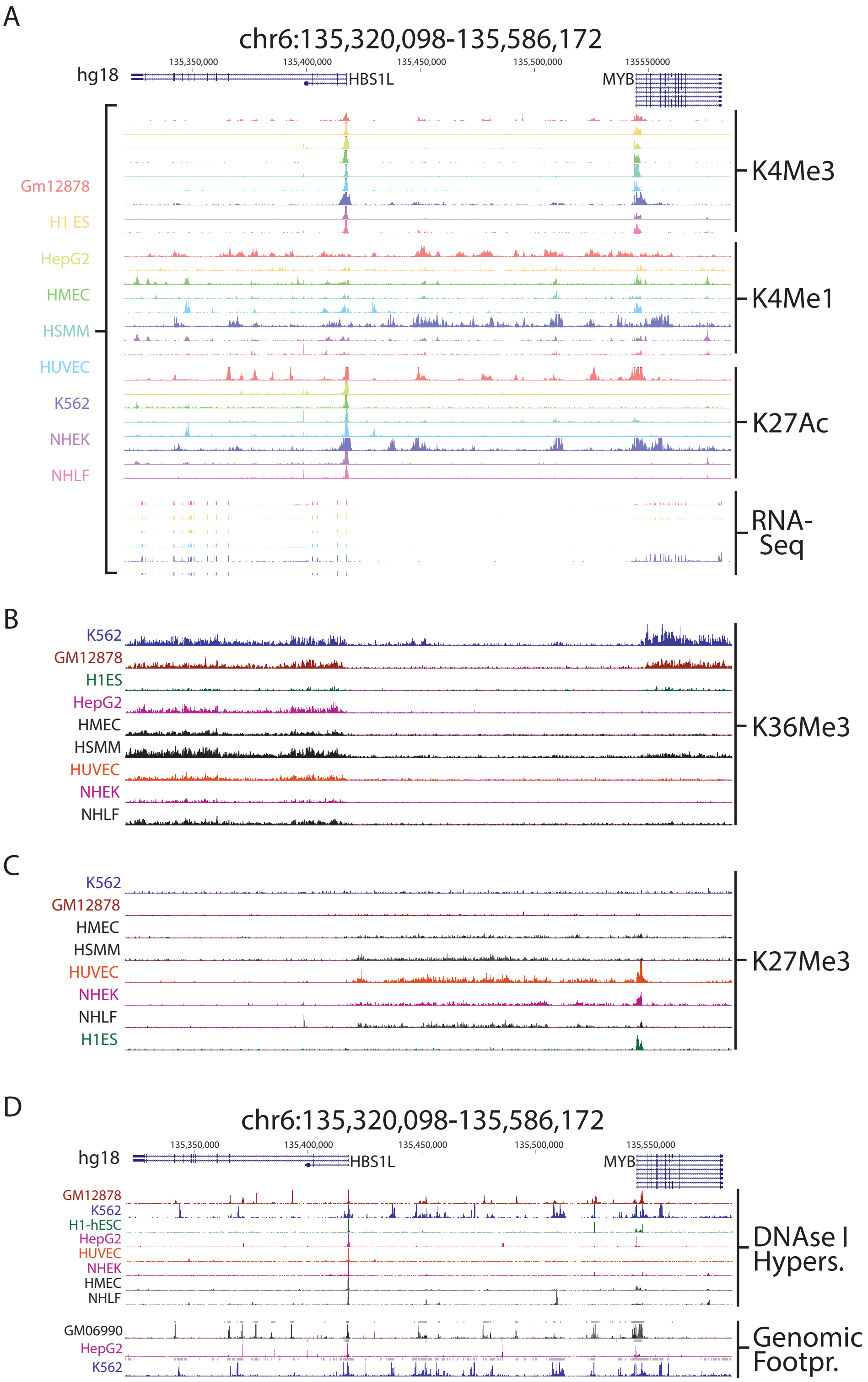
14. Langmead B, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.

15. Norton N, et al. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum Genet*. 2002;110(5):471-478.

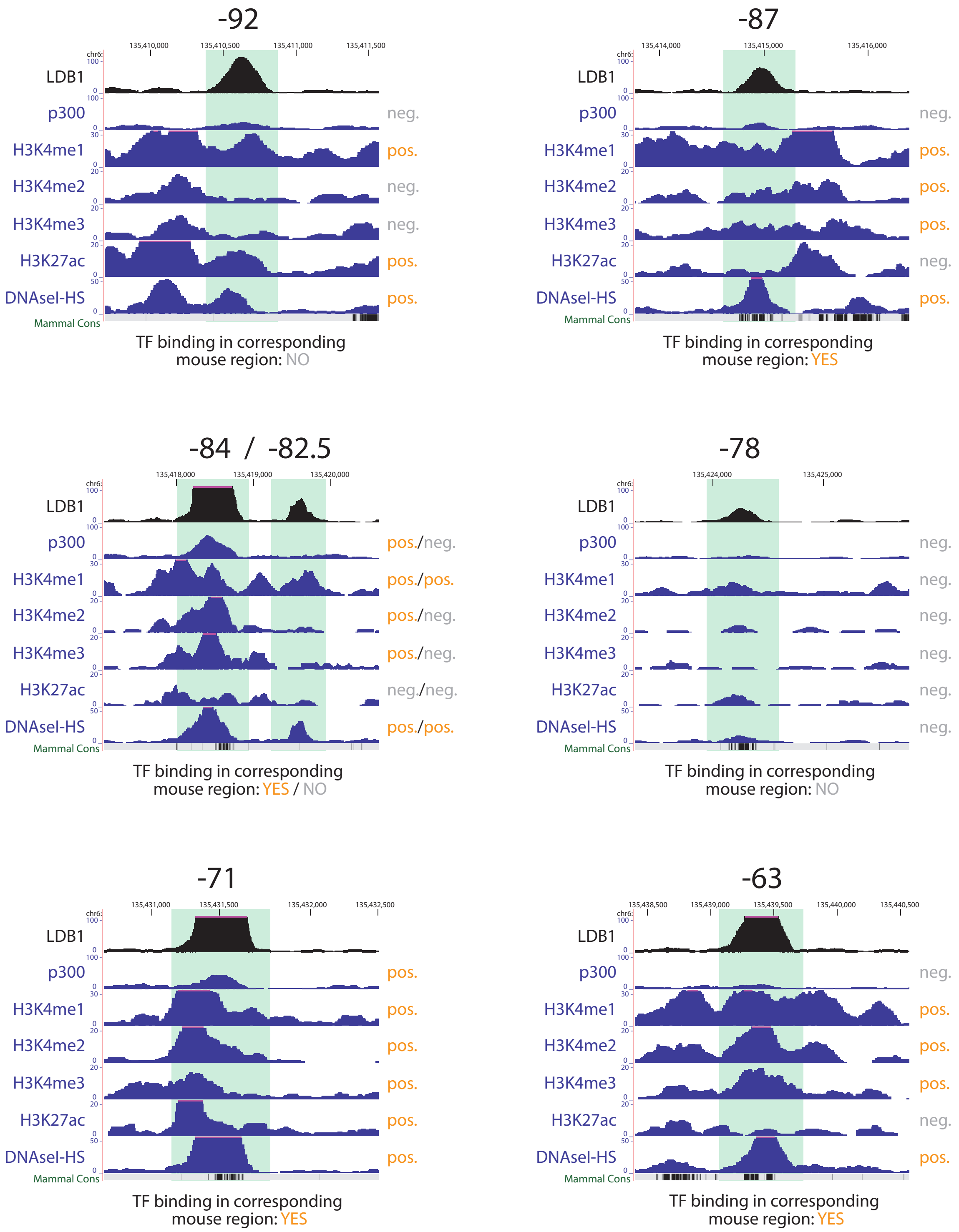
Stadhouders, Aktuna et al. - Supplementary Table 1

primers used in this study:

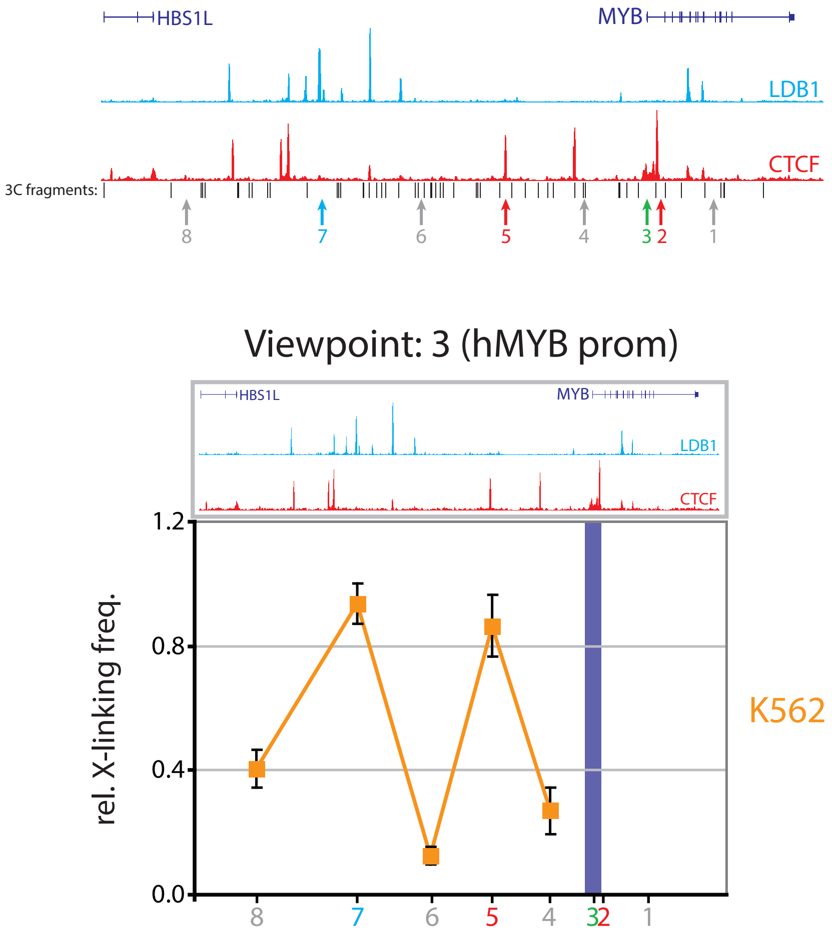
Reaction	Application	Primer 1	Primer 2
hLDB1 -92	ChIP-qPCR	ACTCCTGAGATTCTACCATTAGG	GTGAAGGTGTTGGCTGATAAA
hLDB1 -87		GACAAACACTCCCACAGG	TGAACCTCACAAAACCCTTG
hLDB1 -84		TTCTTCTCCTCTATGGTAACCG	TTATTTTCACAACCGAGCCT
hLDB1 -82.5		GCAACCACACCCGACTAA	ACACCTGTAATCATGCTTTAGG
hLDB1 -78		GCCTGCTGTAGACTGCTA	GCTGACACCCACTTCAAA
hLDB1 -71		GTTACTGAAGAAGGCTGCTG	ACTGAGAGACCAAAGACACT
hLDB1 -63		GAACAACCTGTCACATAAGAC	TGAGACGGGAAGGAACTTT
HBG1/2 prom		AAACGGTCCCTGGCTAAACT	GCTGAAGGGTGCTTCCTTTT
hAMY prom		GGCTGAGTGTTCTGGGAT	CACGGTGCTCTGGTAGAT
alpha HS40		CGACCCTCTGGAACCTATCAG	TTGGCCTCCAGAAGCACTG
MYB	Gene expression	CAACGACTATTCCTATTACCACA	CTAAACTAACCGTCTCTCTCAT
HBS1L		AGTATCACCACTTGTAAAGAAAG	GAGCCCGAATCCTCATCT
ACTB		tgacccagatcatgtttgag	cgtacagggatagcacag
LDB1		TAGCACCTTCGCCCTCTC	TTGGCTGCGTCAAACTGG
TAL1		AAGGTCTCCTCTTCACTCG	CTCTACAGCCTCAGCCAG
KLF1		CCCTCCATCAGCACACTG	TGGTCCTCAGACTTCACG
MYB prom + -84	3C-qPCR	ACCACACCGGGTCTAGCA	CCATTAAACAGTAACTCCCCAGAT
MYB prom + -16		ACCACACCGGGTCTAGCA	TGGGATGCTACCTTCAAAATCAAG
MYB prom + -36		ACCACACCGGGTCTAGCA	CAATGCATTGATCCTCTGGTAGAA
MYB prom + -53.5		ACCACACCGGGTCTAGCA	TGTGAGTGCGAGAACCCAAAG
MYB prom + -4.3		ACCACACCGGGTCTAGCA	TCAGATGTATCACAGTGCTTAGAT
-84 + CTCFintr1		CCATTAAACAGTAACTCCCCAGAT	TCAACTAAACTACAGGTGCTCAAG
-84 + -36		CCATTAAACAGTAACTCCCCAGAT	CAATGCATTGATCCTCTGGTAGAA
-84 + -16		CCATTAAACAGTAACTCCCCAGAT	TGGGATGCTACCTTCAAAATCAAG
-84 + -4.3		CCATTAAACAGTAACTCCCCAGAT	TCAGATGTATCACAGTGCTTAGAT
-84 + -53.5		CCATTAAACAGTAACTCCCCAGAT	TGTGAGTGCGAGAACCCAAAG
-84 + +19		CCATTAAACAGTAACTCCCCAGAT	CAGCAACAATTCCAGATTCATCAG
XPB		CCAGCATGGAGTAGGTGCTAA	TGGAGCCACCATGGGATGA
CTCFctrl		TTTCCATGACATGCTCCTCTTG	CCGGCTGATGTTCCAGTTCA
3C-seq MYB prom	3C-Seq	ATTAGCCTAATAAAATTACTGCATG	GACAACTTTACACATTTCATCAC
3C-Seq -84		CTGTACCCTCAAAACACTTCA	TACTCTGTCTCTAATAGGAGATCA
rs66650371 all.sp. quant	Allele-specific ChIP/3C	TTCTTCTCCTCTATGGTAACCG	TTATTTTCACAACCGAGCCT
3C MYB prom + -84 all.sp.		ACCACACCGGGTCTAGCA	GTGATGTTGCAGAAGAACAAAGC
3C MYB prom + -84 all.sp. CTRL		GTGATGTTGCAGAAGAACAAAGC	ATGGTTAATGATCATGGCTGCAA
-84 GATA1 ChIP		TGGTTTTAGGCTCGGTTGTGAAAA	GTAAGTGTCTTCTGAGGGAACCG
-71 GATA1 ChIP		AGGCTGCTGGCTTCTTTGCT	AAGATATCACACCCACCATGTTGA
extension primer ChIP		ATTCGATTCTACTACTGACA	
rs210796 RNA		accgtatttctgtacaagctcta	attatgagtcacacaccaatgaaga
extension primer RNA		TTTTCCTATTTGAGGAAGCAGGTAAA	



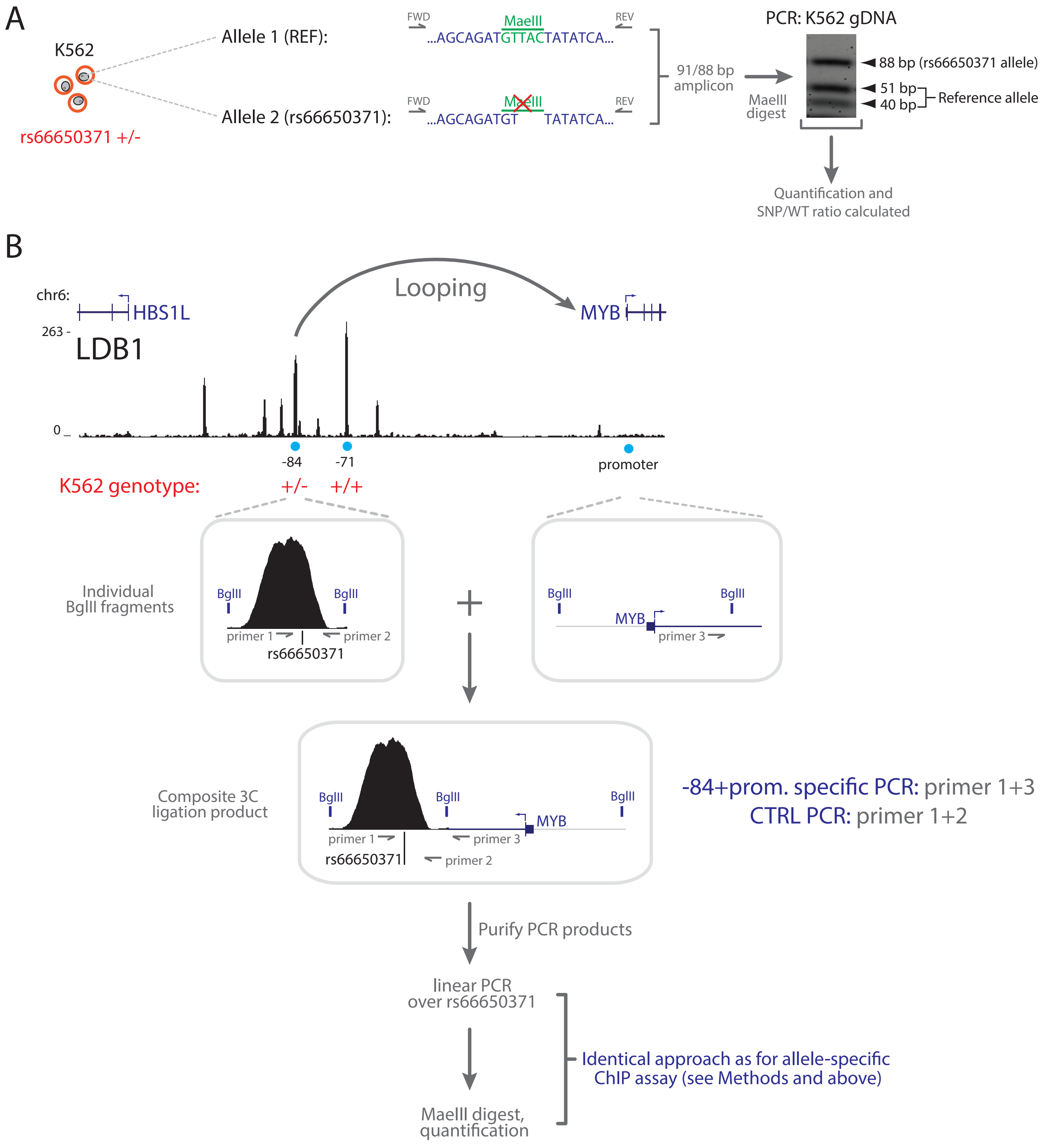
Supplementary Figure 1: Genome-wide histone modification, expression and DNaseI hypersensitivity analysis reveals an erythroid/haematopoietic-specific regulatory signature for the human *HBS1L-MYB* intergenic region. (A) Genome-wide ChIP-Seq and RNA-Seq data from the ENCODE consortium is displayed for the human *HBS1L-MYB* intergenic region in 9 different human cell types. Histone 3 lysine 4 trimethylation (K4Me3, marking promoters), lysine 4 monomethylation (K4Me1, marking enhancers), lysine 27 acetylation (K27Ac, marking enhancers) and RNA-Seq expression analysis are shown. **(B)** Histone 3 lysine 36 trimethylation (K36Me3, marking productive transcription elongation) and **(C)** lysine 27 trimethylation (K27Me3, marking Polycomb-repressed regions) ChIP-Seq data for the HBS1L-MYB intergenic region in 8-9 different human cell types. **(D)** DNaseI-Seq and Digital Genomic Footprinting data for the HBS1L-MYB intergenic region in different human cell types.



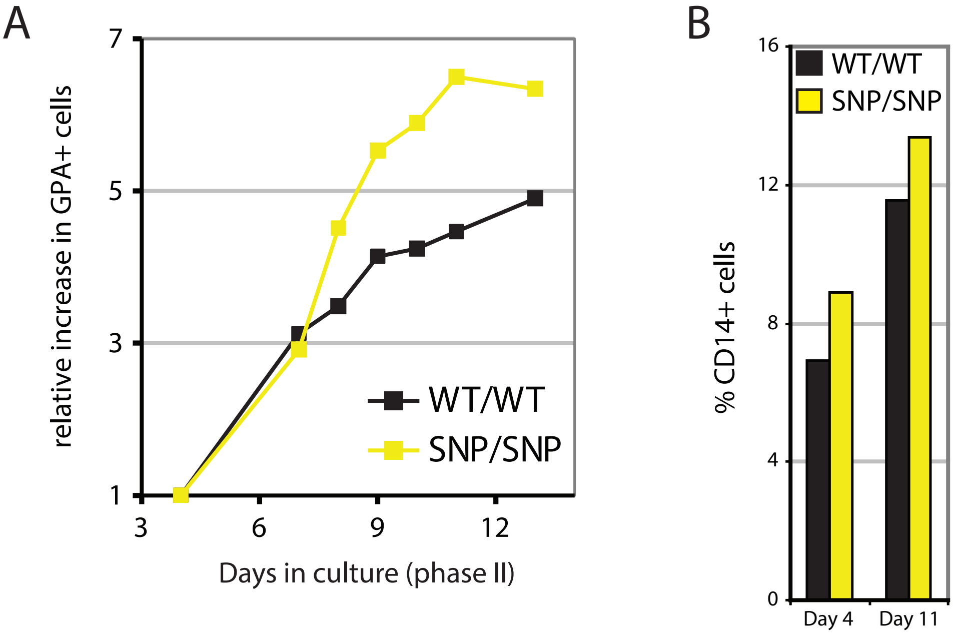
Supplementary Figure 2: Colocalization of DNaseI-hypersensitivity, conservation and enhancer-associated histone modifications and proteins with intergenic LDB1-complex binding sites. Colocalization (highlighted by blue shading) of the different enhancer-associated marks (K562 tracks in blue, obtained from the ENCODE consortium) and the individual intergenic LDB1-complex binding sites (numbered by distance to the *MYB* transcription start site, LDB1 ChIP-Seq track in black) from primary human erythroid progenitors. Mammalian conservation (Mammal Cons) is shown in the bottom track. Transcription factor (Ldb1-complex) binding in the corresponding mouse region (Ref.25) is denoted below each graph.



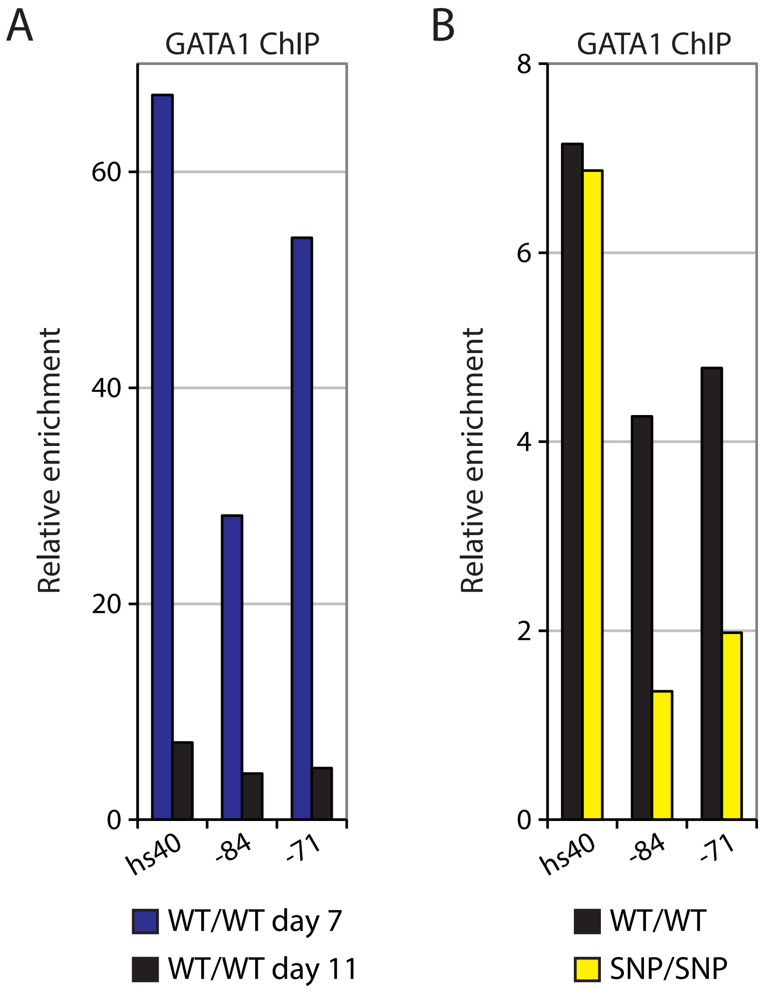
Supplementary Figure 3: Chromosome conformation capture analysis of the *HBS1L-MYB* locus reveals long-range interactions between intergenic elements and the *MYB* gene in K562 cells. 3C-qPCR experiments on K562 cells (n=4) using the *MYB* promoter as viewpoint. The locus is plotted on top, with the different 3C restriction fragments (BglII) used indicated. Interaction frequencies between 2 fragments within the *ERCC3* locus were used for normalization. Error bars display s.e.m.



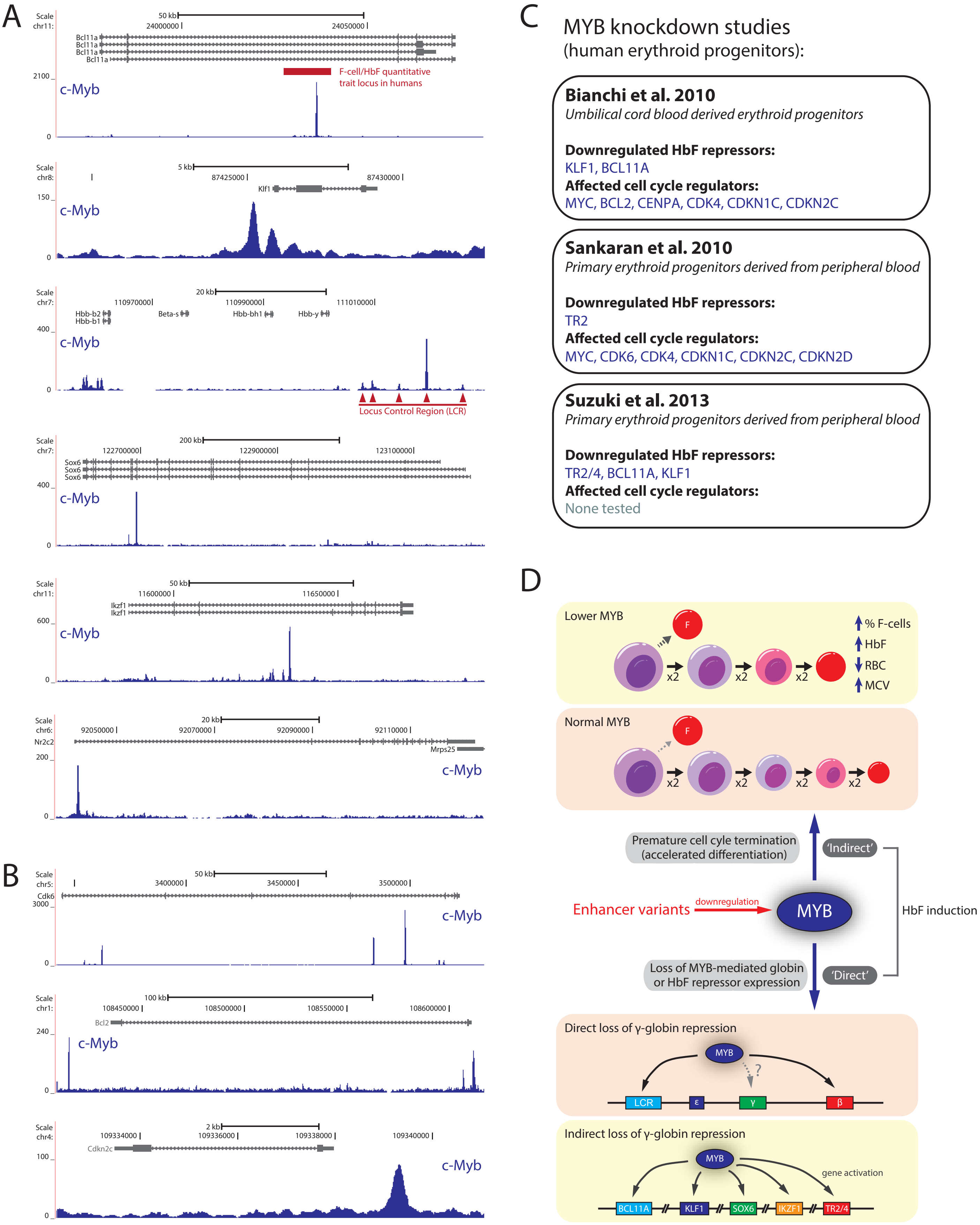
Supplementary Figure 4: Strategy used to quantify differences in transcription factor binding and promoter looping to the K562 rs66650371 alleles. (A) *Mae*III digestion read-out method used for allele-specific ChIP and 3C assays. The rs66650371 reference allele contains a *Mae*III site, which is removed by the rs66650371 3bp-deletion allele. A 91bp product spanning rs66650371 was PCR amplified and subjected to *Mae*III digest, resulting in 3 fragments: 88bp (representing the 'SNP allele') and 51/40 bp fragments (together representing the 'WT allele'). Fragments were separated using agarose gel electrophoresis and quantified to determine a SNP/WT ratio. Duplicate samples spiked with pGL3 plasmid were digested in parallel to ensure digestion efficiencies were similar across samples. **(B)** Strategy used for rs66650371 allele-specific 3C in K562 cells. Below a schematic of the locus (top), 2 boxed figures depict the 2 *Bgl*II restriction fragments that form the composite -84/promoter 3C fragment. Three primers were designed that generate 2 PCR amplicons (~4.5kb) encompassing rs66650371. Primers 1+3 can only yield a product when the specific -84/promoter composite fragment is present ('-84+prom. specific PCR'). Primers 1+2 are both located on the -84 fragment and will amplify all -84 fragments ('CTRL PCR'). Amplicons are purified from gel and the *Mae*III digestion read-out approach described in (A) is used to determine allelic ratios. Ratios obtained from the 'CTRL PCR' were used for normalization.



Supplementary Figure 5: Differentiation kinetics of WT/WT and SNP/SNP cultures as assayed by FACS analysis. (A) Representative FACS measurements showing the relative increase of GPA positive (late erythroid cells) cells during phase II culture (day 4-day 13) of primary erythroid progenitors. The percentage GPA positive cells was normalized intra-individually for small differences in the percentage of CD71 positive cells; day 4 measurements were set to 1. Cells were obtained from individuals homozygous for the minor allele of the phenotype-associated SNPs (HMIP-2 LD block variants; SNP/SNP) and wildtype control individuals (WT/WT). **(B)** Percentage of CD14+ monocytes present in WT/WT and SNP/SNP erythroid cells during culture. Representative measurements performed at day 4 and day 11 are depicted.



Supplementary Figure 6: GATA1 ChIP experiments on primary human erythroid cultures harvested at terminal stages of differentiation. (A) Representative results of ChIP-qPCR experiments for GATA1 on chromatin prepared from WT/WT HEPs on day 7 and day 11 of erythroid differentiation. **(B)** Representative results of ChIP-qPCR experiments for GATA1 on chromatin prepared from WT/WT and SNP/SNP (homozygous for the minor allele of the phenotype-associated HMIP-2 LD block variants) HEPs on day 11 of erythroid differentiation. The -84 and -71 regulatory elements were assayed for GATA1 binding, the α-globin hs40 region was used as a positive control. Enrichments were normalized to IgG.



MYB knockdown studies
(human erythroid progenitors):

Bianchi et al. 2010
Umbilical cord blood derived erythroid progenitors

Downregulated HbF repressors:
KLF1, BCL11A
Affected cell cycle regulators:
MYC, BCL2, CENPA, CDK4, CDKN1C, CDKN2C

Sankaran et al. 2010
Primary erythroid progenitors derived from peripheral blood

Downregulated HbF repressors:
TR2
Affected cell cycle regulators:
MYC, CDK6, CDK4, CDKN1C, CDKN2C, CDKN2D

Suzuki et al. 2013
Primary erythroid progenitors derived from peripheral blood

Downregulated HbF repressors:
TR2/4, BCL11A, KLF1
Affected cell cycle regulators:
None tested

Supplementary Figure 7: Plausible model for *MYB*-mediated repression of HbF levels via cell cycle regulation and transcriptional activation of HbF repressor genes. (A) c-Myb ChIP-seq data (obtained from the ENCODE consortium) from MEL cells showing c-Myb binding to the β-globin locus and HbF repressor genes (*Nr2c2* encodes the TR4 protein). **(B)** c-Myb ChIP-seq data from MEL cells showing c-Myb binding to selected cell cycle regulators. **(C)** Analysis of published MYB knockdown studies in human erythroid progenitors. Downregulated HbF repressor genes and a selection of affected cell cycle regulators is shown. **(D)** Dual model of *MYB*-mediated HbF repression. Lower *MYB* levels (as a result of disrupting enhancer variants) can lead to HbF induction via increased premature cell cycle termination ('indirect', top part), resulting in the generation of more F-cells and a higher HbF level. Fewer proliferation cycles ('x2', indicating cell division) will result in a lower red blood cell count (RBC) and a larger mean cell volume (MCV). Alternatively, lower *MYB* levels could result in a loss of proper transcriptional regulation at the β-globin locus and HbF repressor genes ('direct', lower part). Reduced activation by *MYB* of known HbF repressors (e.g. BCL11A, KLF1) or disrupted regulation at the β-globin locus could result in γ-globin gene reactivation and subsequent HbF induction.